THE ETHICS AND EPISTEMOLOGY OF DEEPFAKES

*Taylor Matthews and Ian James Kidd*

Forthcoming in Carl Fox and Joe Saunders (eds.) *The Routledge Handbook of Philosophy and Media Ethics*. Please cite published version.

## 27.1 Deepfakes: A Primer

In popular usage*, deepfakes* are often understood as ultra-realistic videos capable of depicting people – usually celebrities or politicians – doing or saying things they never did or representing a state of affairs that never obtained. Occasionally, this term is mistakenly used interchangeably with so-called *shallowfakes*. These are videos that are slowed down, sped up, or whose audio is tampered with, all to depict people doing or saying things they never did. Photoshopped videos are a good example of this. A deepfake, by contrast, can be defined broadly as “any mimicry, manipulation, or synthesis of video or audio that is enabled by machine learning” (Dhillon, 2019). A video does not count as a deepfake unless it relies on deep-learning (machine) artificial intelligence AI to generate the content. Well-known examples include deepfakes of former US Presidents Barack Obama and Donald Trump and a TikTok account, @deeptomcruise, devoted to deepfakes of actor Tom Cruise. Some are funny, some are manipulative, some are used to inform and educate, and they are developing so rapidly and have such significant implications that they have been described as “one of the most worrying fruits of rapid advances in artificial intelligence” (Galson, 2020). Against dramatic claims that the era of deepfakes constitutes an “epistemic apocalypse”, some philosophers urge a calmer and soberer perspective ((Habgoode-Coote, 2023, Harris, 2021). Still, this leaves ethicists, technologists, policymakers and others with plenty to think about, and this chapter surveys some of these moral and epistemological issues.

“Deepfake” is a portmanteau of *deep learning* and *fake*. The deep learning involved in the creation of a deepfake comes in the form of a deep *neural network*, which can be roughly understood as a non-linear model for predicting and generating particular content – usually video content – based on the input of images and video footage (Mirsky and Lee, 2020: 5–6; Westerlund, 2019). The neural networks typically responsible for deepfakes are so-called Generative Adversarial Networks, or GANs, consisting of two algorithms called the *generator* and the *discriminator* (Mirsky and Lee, 2020). After being fed the same dataset of audio, video footage, and images of one person or target, the adversarial component kicks in and the two algorithms “compete” against each other: the generator creates new samples based on the initial input that are good enough to trick the discriminator, while the discriminator

DOI: 10.4324/9781003134749-33

tries to determine if this new output is real. Eventually, the GAN becomes capable of manufacturing a video capable of realistically mimicking the voice, mannerisms, facial expressions, and speech patterns of one person then superimposing them onto another – hence the “fake” part of “deepfake” (Mirsky and Lee, 2020).

We can distinguish two categories of deep-learning facial manipulation (Tolosana et al., 2020: 3). First, synthesising photographs and video input to generate entirely non-existent faces (*entire-face synthesis*). Second, manipulating particular facial attributes that might be characteristic of certain people, which ranges from adding or removing features such as moles and wrinkles to more extensive changes like adjusting the skin colour of a person’s face (*attribute manipulation*). Third, deep-learning can also swap or replace the facial expressions of one person with those of another, like smiles, snarls, grimaces, looks of confusion or disgust and so on (*expression swap*). Fourth, deep-learning techniques can replace the identity of one person and swap it with another (*identity swap*). Unsurprisingly, this has become the hallmark of deepfakes.

Considering these versatile functions, the deep-learning technology used in deepfakes has a range of applications in a range of creative industries. Within the video-game and film industries, entire-face syntheses can be used to manufacture a character’s face from existing images and videos of people, a famous case being the recreation of the late Peter Cushing as the character Grand Moff Tarkin in the *Star Wars: Rogue One* film. Opinion was divided about the “digital resurrection” of deceased actors; one commentator judged the result as creepy, “a dead man’s face, a living man’s voice, a whole team of programmers’ code” (Sargeant, 2017: 32). The technology can also be used to create convincing fake online profiles for social media platforms, with source material often drawing on websites like thispersondoesnotexist.com, which uses deep learning to generate completely unique faces of people who do not exist. Moreover, attribute manipulation – also called *face-editing* or, more innocuously, *retouching* – allows apps such as Instagram and Snapchat to create filters over users’ faces, add or remove features such as glasses, moles, or wrinkles, and even lighten or darken one’s skin colour. Expression swap has been used to insert smiles or grimaces into otherwise stern videos or photos; one example is a video depicting Winston Churchill and Adolf Hitler singing.1 Less incongruously, identity-swap technology has been used in advertising to dub the footballer, David Beckham, speaking in a variety of languages, raising awareness of the dangers of malaria (Roettgers, 2019).

A final example of the possibilities of deepfake technology involves the development by researchers at Adobe and Princeton University of so-called VoCo technology. This allows creators to alter the content of an audio recording by typing the alternative desired speech into a transcript (Jin et al., 2017; Rini, 2020).2 By feeding voice samples into the VoCo algorithm, one can analyse and synthesise what the target’s voice *would* sound like *were* they to say the things typed into the transcript. In practice, deepfakes developed using this technology are now able to depict people saying *anything* at the behest of the creator, where the limits are moral and prudential rather than technological in character. Importantly, this ability marks a watershed moment in current deepfake technology. Whereas deepfakes were often given away by the artificial mapping of facial movements onto the target speaker, VoCo technology now allows purveyors to generate a “voice” to accompany a deepfake. Doing so improves the perceived authenticity of deepfakes and the credibility with which viewers assign such videos.

It should be clear from this overview that deepfake technologies and their applications raise a number of ethical, political, and epistemologically charged issues. Our aim in this chapter is to focus on several issues that, while geared towards epistemology, also have important upshots for those working within media ethics. These include issues of trust, our dependence on others, the role of luck in the formation of our beliefs, and the integrity of our ethical and epistemic rights within a world of deepfakes.

## 27.2 Deepfakes, Information, and Trust

Deepfakes are an active topic of academic research: between 2018 and 2020, the number of articles published on this subject rose from three to over 250 (Mirsky and Lee, 2020). Only recently, however, have philosophers begun investigating this new form of media manipulation. Initially, much of this work has focused on the legality of deepfakes (Chesney and Citron, 2019; Gieseke, 2020; Kirchengast, 2020; Pfefferkorn, 2020), or their ethical ramifications (Chesney and Citron, 2019; Diakopoulos and Johnson 2021; de Ruiter, 2021; Harris, 2021; Öhman, 2022, and Rini and Cohen, 2022). For this reason, we use this chapter to shed light on a number of epistemologically oriented worries that deepfakes might generate. In part, we do so because a welcome feature of epistemology is its increasing attention to roles of material and technological infrastructure in epistemic life and to technologically mediated phenomena like fake news (Lynch, 2016; Rini, 2017; Gelfert, 2018).3 However, we also do so because of the often inseparable ethical and epistemic dimensions of our lives. By surveying the epistemic challenges posed by deepfakes, then, we naturally unearth wider ethical ramifications, and vice versa.

### *27.2.1 Evidence and Information Sharing*

In an online world increasingly populated by “digital natives”, videos are an invaluable source of perceptual evidence about people and states of affairs from within our personal networks and our local, national, and international communities. First-hand perceptual evidence, by its nature, is limited: we can’t be everywhere, or reliably record unfolding events using our native biological capacities. In these cases, videos offer an alternative, especially for events that are beyond our reach, like historically and geographically distant states of affairs. Don Fallis (2020) recognises that videos meet these important informational needs and worries that deepfakes will prevent people from acquiring knowledge through a number of means. First, if we cannot detect that we are watching a deepfake, we are likely to acquire *false beliefs* about a person or state of affairs. If we fail to detect the deepfake, then we also fail to identify true beliefs about the sources of our beliefs. Second, deepfakes could undermine our confidence in the reliability of our *belief-forming processes*. The sorts of evil-demon cases that provoke deep doubts about our perceptual and doxastic functions may only occur under certain rather rare conditions; deepfakes, however, could be ubiquitous within our lives. Indeed, they could lead to consistent hesitation or suspension of judgement about what we watch, putting us in a situation of persistent scepticism about a socially entrenched informational resource (Fallis, 2020: 3). A third concern about deepfakes, related to the other two, is that we may no longer be justified in believing the testimony we hear from videos – or, more complexly, even if we *are* justified in accepting the testimony of deepfakes, we will find it difficult to ascertain this for ourselves. How one articulates this worry depends on what sort of epistemology of testimony you want to adopt. According to one influential account, we’re justified in trusting testimony if and to the extent that it is well-supported by the evidence. A friend might tell us they’ve got their head shaved, then sent a TikTok video of the shaving. But if we can no longer trust video evidence, it will become more difficult to trust at least certain testimonies (Leonard, 2021).

The primary concern for Fallis, however, is the extent to which deepfakes could decrease the amount of information that videos provide to their viewers. This can happen in lots of ways: a video may depict less detail of a person or an event or be purposefully vague in its depictions or be edited in problematic ways. Fallis is concerned with the evidence videos impart and draws on Bryan Skyrms (2010) account of information sharing. According to Fallis, the amount of information a video carries depends on the probability of the video existing as opposed to the probability of it not existing. The lower this probability, the more information the video carries (2020: 9). To borrow his example, a video of a corrupt politician taking a bribe carries more information or evidence about the politician’s actions if the probability of this video existing is higher than the probability of it not existing (2020: 9). The worry with deepfakes is that they will become increasingly sophisticated, making it increasingly difficult to distinguish them from genuine videos. In such situations, one could work increasingly hard to ascertain their status, or one could suspend judgment, or one could seek alternative sources of information, whatever they may be. Deepfakes will become an increasingly prevalent feature of our informational ecosystem, even becoming a dominant presence. By changing our informational environment, they necessitate changes in personal epistemic behaviour. A main area of change will be the new challenge of *detecting* deepfakes, of becoming able to distinguish them from genuine videos if possible, and of adapting doxastic behaviour if not (Lang, 2019). One key method in the detection of deepfakes, or *digital forensics*, is to draw on GANs or other AI. But the problem with detection systems relying on GANs is that the AI used to power them could be turned against them to bypass initial detection (Farid, 2018: 268). Thus, the main worry is that we cease to be able to detect them; if the technology develops at its current pace, then without some corresponding advances in technologically augmented detection, we lose the race. Eventually, we may not be able to tell when we are watching a deepfake, and so become vulnerable to acquiring false beliefs from them if we choose to trust their contents.

In this situation, two worries become salient: if deepfakes decrease the evidential weight of videos by increasing the probability of false positives, they risk us accumulating more false beliefs and, more subtly, impact on the number of true beliefs we *ought* to gain. I may cease to *expect* that I’ll gain some or many or any true beliefs from watching online video content. In an informational world where video content is a main source of information, that is a disturbing prospect. Of course, one can decide to suspend judgement about information gained from videos because one recognises it may be deepfaked (Fallis, 2020: 11). But this may have the problematic effect of undermining our ability and willingness to judge correctly and believe truly – a counsel of despair, if ever there was one.

To explore that dire possibility, let’s consider some different ways to think about how deepfakes might impact our epistemic lives.

### *27.2.2 Epistemic Trust and Backstop Crises*

Fallis frames the epistemic threats of deepfakes in terms of a capacity to undermine evidence and information-sharing. A different account is offered by Regina Rini (2020), who argues that the threat they pose is not so much about people falsely believing deepfakes, but rather about their capacity to “gradually eliminate the epistemic credentials of *all* recordings” (2020: 8). The worry starts with the observation that video recordings play a pervasive, albeit implicit, role: they help to correct and regulate our testimonial practices. Think of those prone to make outlandish claims or deny actions attributable to them: one way we establish the truth about these cases is by appealing to a video recording – a common strategy of exposing hypocrisy and bullshitting used by satirists, investigative journalists, and whistle-blowers. In these cases, an electronic public record serves two valuable social-epistemic function: first, it helps us to correct or call out particular claims and, second, it can help to motivate us to regulate personal epistemic conduct by creating the possibility of our current claims being evidentially tested. I cannot be sure I’ll get away with exclaiming, “I never said that!”, if I’m aware that there may be a video recording of my saying just that somewhere. In these cases, video recordings can instrumentally encourage us to be truthful or sincere testifiers and help us navigate in whom we can and cannot trust. Rini calls this the *epistemic backstop*.

With the advent of deepfakes, however, Rini fears we will enter a *backstop crisis* – a time when video recordings no longer act as a corrective and regulatory tool for our testimonial practices. Suppose a malicious actor creates a sophisticated deepfake of President Biden rigging the US Electoral College. With the help of VoCo technology and the latest GANs, a purveyor could easily manufacture a convincing deepfake that looks and sounds like Biden. Naturally, Biden would insist that he did no such thing, but the very presence of the fake would at least give air to Trump supporters who believed the election was rigged: we can, here, distinguish cases where deepfakes exacerbate existing epistemic problems and where they create novel epistemic problems. More worryingly, if a public official were recorded engaging in a criminal act, they could simply “cry deepfake” and downplay the video’s credibility. If deepfakes facilitate this sort of behaviour, we will have less reason to trust what we see on video, knowing that anybody could feasibly testify insincerely and untruthfully with little, if any, ramifications. Given that a number of political leaders already engage in this kind of behaviour, it seems as though deepfakes will only exacerbate this problem.

Over time, Rini thinks that deepfakes will facilitate successive backstop crises, which will transform how we approach evidence from videos. As Rini points out, the evidential weight of videos currently occupies that of perceptual evidence, much like still photographs (Cavedon-Taylor, 2013; Cohen and Meskin, 2004; Hopkins, 2012). This is contrasted with paintings or drawings, which offer us weaker justification akin to testimonial evidence, and it is this stronger justificatory status that allows videos to act as a good backstop for our testimonial practices. Yet, as videos become more susceptible to deepfakery – and, thus, open to wider interpretation about whether they are genuine or not (see Groh et al., 2022) – Rini believes that we will downgrade the status of video evidence from that of perceptual to testimonial evidence. Since testimonial evidence is weaker than perceptual evidence, deepfakes will undermine the trust or credentials we assign to videos, if we can no longer reliably invoke such media to regulate or correct others’ claims and assertions (see Atencia-Linares and Artiga, 2022).

## 27.3 Luck, Rights, and Character

In the previous section, we canvassed two epistemic concerns that philosophers have identified in relation to deepfakes. In this section, we will consider three currently unexplored questions and problems that deepfakes could give rise to. First, given the expanse of deepfakes, what role does *luck* play in the beliefs we form upon watching videos? Second, do deepfakes violate our *epistemic rights*? Third, could deepfakes adversely affect our *epistemic character*?

### *27.3.1 Deepfakes and Epistemic Luck*

In addition to the concerns articulated above, there is a third problem on the table that Fallis gestures at but leaves to one side. In particular, he writes that even if we happen to watch a genuine video in spite of it possibly being a deepfake, our belief “probably does not qualify as knowledge” because our belief is *much less safe* (Fallis, 2020: 10, our emphasis). What Fallis picks up on is the notion of *safety*, roughly according to which one’s true belief counts as knowledge so long as the belief could not too easily have failed to obtain (i.e. turn out true) in close possible worlds, while formed on the same basis (Pritchard, 2005).4 Those who attach a safety condition on knowledge do so in order to combat the presence of what epistemologists call *epistemic luck*. In this section, we will argue that deepfakes can render our beliefs from videos lucky, and, hence, cause us to lose out on knowledge. To begin, though, it will be useful to explore the notion of epistemic luck further. As the term suggests, epistemic luck corresponds to the way in which our beliefs can turn out true or false as a result of luck. Until recently, many epistemologists thought that knowledge and luck were incompatible – at least, it seems odd to say that you know a proposition if your answer was correct but the result of a lucky guess. On closer inspection, though, there are forms of epistemic luck that are compatible with knowledge. Mylan Engel (1992) identifies a form of *evidential* luck as one such kind, where it is just a matter of luck that one has evidence for a belief but, given that evidence, one’s belief is not true by luck. Evidential luck is contrasted with what Engel calls *veritic* epistemic luck, where it is just a matter of luck that one’s belief is true. Veritic epistemic luck, Engel argues, is the kind of luck which is incompatible with knowledge.

Duncan Pritchard (2005, 2015) distinguishes between two types of veritic epistemic luck: *intervening* and *environmental* luck.5 As an example of the former, suppose you are standing in a field and believe you see a sheep in the next field. However, a dark rain cloud overhead reveals that the animal is actually a fluffy white dog. As it turns out, though, there is in fact a sheep hidden behind the dog.6 Here, your belief about the sheep is true *in virtue* of the intervening presence of the sheep. Had it not been for its lucky appearance, your belief would have been false. As an example of the latter, imagine you are driving through the countryside and see dozens of barns on the hillside. Unknown to you, the majority of these barns are barn façades, manufactured to entice tourists to the area. However, the barn you look at is one of the few genuine barns on the hillside (Goldman, citing Ginet, 1976). Unlike the previous case, your belief is true not in virtue of any intervening factors; rather the highly lucky features of your environment – the presence of a genuine barn amongst many barn façades – seems to impinge on your claim to know.7

With these species of epistemic luck clearer, we can turn to consider how they might relate to deepfakes. Since our focus is on the way in which deepfakes might undermine the epistemic status of our beliefs rather than the evidence videos provide, we limit our discussion purely to veritic epistemic luck.8 How, then, might deepfakes generate the sort of intervening and environmental luck that Pritchard identifies? To set the scene, let us draw on events from the world of medicine. In 2019, researchers at a hospital in Israel were tasked with creating deepfake CT scans, which depicted malignant cancer tumours. Concerned about the vulnerability of medical infrastructure to a cyber-attack, the researchers tested whether three radiologists of varying experience were able to identify the CT scans with injected deepfake tumours from genuine CT scans (Mirsky et al. (2019: 3).9 Bearing this situation in mind, we can construct a case of environmental luck. Suppose you are a radiologist at the hospital above and you are tasked with examining a batch of CT scans for malignant tumours. Your wealth of experience allows you to correctly conclude that there are no malignant tumours in the particular CT scan you watch. Thus, you form the true belief that it is healthy. Unbeknownst to you, though, a malicious agent has intercepted your particular batch of CT scans and added deepfake tumours into the others. As it turns out, though, the CT scan you analysed was genuine (Matthews, 2023). Now for an example of intervening luck. Place yourself in the shoes of the same radiologist, tasked with analysing a patient’s CT scan for cancer tumours. Again, drawing on your experience, you analyse this scan and conclude that this one contains malignant tumours. This time, however, the scan is not genuine. Rather, a hacker intercepted your patient’s records and added deepfake tumours into the CT scan. Nevertheless, it turns out that the patient under examination is actually living with cancer and, thus, your belief is true. Given this information, does the radiologist *know* in either case? Intuitively, it seems not. In the first case, the deepfake scans replicate the barn façades in the example above, and so, while you arrive at your true belief through exercising your cognitive abilities, you very easily could have selected another CT scan and formed a false belief. The lucky nature of your environment as a radiologist – that you selected a genuine scan amongst the deepfakes – causes us to hesitate ascribing knowledge despite the cognitive success. Similarly, in the second case, the presence of the deepfake scan acts as source of bad epistemic luck, yet, the patient’s actual cancer diagnosis intervenes to cancel this out, affording the radiologist with a true belief. Despite the differences in the way luck plays out, both cases illustrate how deepfakes could well introduce veritic epistemic luck into our lives, and with it undermine our claims to knowledge from videos. One might initially question whether the sort of veritic luck introduced by the deepfakes above could *really* feature in the videos we watch more generally. After all, deepfake tumours are surely harder to detect than deepfakes of facial or bodily movements. Nevertheless, the development of real-time GANs and VoCo algorithms has allowed purveyors to manufacture highly sophisticated deepfakes that are increasingly indiscernible from genuine videos. A notable example here was the deepfake of Queen

Elizabeth II, created by the UK television broadcaster Channel 4 (Channel 4, 2020).10

Second, with easier access to such technology, it is no surprise that the number of deepfakes in circulation has increased. Estimates from 2019 put this figure at 14,678 (Adjer et al., 2019), but it is almost certain that this number has risen just as the versatility and credibility of deepfakes has increased. Still, one might be inclined to think that we can check the authenticity of a deepfake by recourse to its genuine source material. As Fallis and Rini’s concerns highlighted, however, this is by no means an easy feat. For starters, the increasing prevalence of sophisticated deepfakes – and, hence, the chance of watching a “false positive” – could diminish the evidence carried by the genuine video if it directly contradicts an apparently authentic video. Subsequently, we might find ourselves suspending judgement about genuine videos rather than using them to verify deepfakes. *Pace* Rini, even if we have justifiable grounds to believe a genuine video, we might nevertheless fail to believe its content because deepfakes undermine our trust in what we see and hear from them. As deepfakes expand into our online environments, then, the risk of veritic epistemic luck – both its intervening and environmental manifestation – entering into our epistemic evaluations simultaneously increases. With the technology developing each year, and the GANs used to power deepfakes evolving to avoid detection (Farid, 2018), deepfakes could bring into question whether we know what we are watching from a video is true, as opposed to a lucky belief.

### *27.3.2 Deepfakes and Epistemic Rights*

Internet ethics researchers have explored the connections between deepfakes and various kinds of rights. Worries have arisen, for instance, that deepfakes could threaten violations of personal data protection rights and intellectual property rights. and the World Intellectual Property Organisation has raised the question of the legal ownership of deepfaked videos (Çolak, 2021; WIPO, 2019). The focus of those discussions are legal and moral rights, but we should also be open to the idea that deepfakes pose challenges for our *epistemic rights*, a concept developed by Lani Watson (2021):

I define an epistemic right as a complex entitlement that provides justification for the performance and prohibition of actions and omissions concerning epistemic goods […] a right is an epistemic right insofar as it concerns epistemic goods [like] knowledge, truth, belief, justification, understanding, wisdom, information, misinformation and ignorance.

#### (Watson, 2021: 15)

To unpack this claim: we have rights that mandate or require actions and omissions that relate to epistemic goods. We have a right to be kept informed about certain things. We have a right not to be deceived by others except under specific conditions. We have a right to know and be told the truth about important aspects of our history, life, or the world. We have a right to privacy and confidentiality, which should stop people sharing personal information about us without our consent. A sad feature of life is that our epistemic rights are often ignored or violated, and Watson notes this is often due to our media environment. “One need only pick up a newspaper or turn to the various social media platforms”, she remarks, “to find accusations of lies, deceit, fake news, and ‘truth-twisting’” and, unfortunately, “in the fast-paced, crowded and information-centric world of the twenty-first century, they are perhaps more pernicious and more damaging than ever before” (Watson, 2021: 53, 89).

Deepfakes plausibly jeopardise our epistemic rights in several different ways, but since there is as yet, to our knowledge, no literature connecting the two, we will simply offer some possibilities. An obvious one is that deepfakes violate our right not to be deceived, especially when being deceived would cause or threaten harm to our interests. By deceiving others with a deepfake, one arguably causes them what Watson calls an *epistemic insult*: a harm done to a person’s status and capacities as an epistemic agent (Watson, 2021: 74). The possibility is that deliberately circulating deepfakes, in full knowledge that you’re doing so, interferes with the good-faith epistemic efforts of other people, constitutes an epistemic insult.

Another possibility is that deepfakes violate our right to know, because what they depict is typically, if not always, false or misleading, like a liberal politician endorsing radically conservative social policies. Sometimes it may be possible to discern this, but not always and deepfakes are typically designed to be test-resistant. I have a right to know things about the world – its history, economics, current affairs, and so on – and much of that knowledge comes from online video content. When that right to know is violated, for instance through our being deceived, one is “put at an epistemic disadvantage that reflects dysfunction in the epistemic system”, and this is what Watson calls an *epistemic injury* (Watson, 2021: 71). Like a physically injured person, the epistemically injured person will find it harder to perform certain actions and pursue certain projects that matter to them, like the everyday epistemic activity of trying to stay more-or-less up-to-date with what’s going on in the world.

It would be worthwhile investigating whether the concept of epistemic rights could be used to articulate the harms of deepfakes. Certainly, there are moral and legal rights relevant to deepfakes, too, but it seems right to say that the epistemic dimensions of deepfakes merit attention, too.

### *27.3.3 Deepfakes and Epistemic Character*

A final potential set of issues for considering deepfakes involve *epistemic character*. The idea of character epistemology is that human beings have epistemic characters, consisting of an array of epistemic dispositions – dispositions to argue, doubt, interpret, investigate, reason, and think in certain ways. Sometimes, these dispositions take the strong and stable forms we call epistemic virtues and epistemic vices, which are studied by virtue and vice epistemology, respectively (Baehr, 2011; Battaly, 2016; Zagzebski, 1996). Character epistemologists explore the nature and significance of epistemic virtues and vices, and this increasingly means careful investigation of how epistemic virtues and vices play out in specific contexts, institutions, and structures (Turri, Alfano, and Greco, 2019; Kidd, Battaly, and Cassam, 2020). Since the epistemic lives of many people in the world are now mediated by a technologically constituted social-epistemic structure that includes social media and video content, we should ask how epistemic virtues and vices are related to that structure and, therefore, to the deepfakes that are now a part of it.

As with epistemic rights, there appears to be limited literature that applies resources of character epistemology to deepfakes. One recent exception is Matthews (2022), who claims that deepfakes have the propensity to be epistemically corrupting, that is, to encourage the development and manifestation of epistemic vices in people. Of particular interest to him is what he calls *intellectual cynicism*, which in its milder form involves a failure to attend to the intellectual merits of online videos as a genuine epistemic source, and in its stronger form involves making inferential leaps about the epistemic credentials of videos more generally. As this brief overview suggests, a natural tendency might be to view deepfakes as problematic from a character-epistemic perspective. But there remain several questions in need of answers, including: Do deepfakes represent or reflect epistemic vices? For instance, if the intention of a deepfaker is to disrupt the epistemic functioning of others, they may be guilty of epistemic malevolence (Baehr, 2010). We can also ask how a person may respond to the knowledge that deepfakes are a feature of their social-epistemic environment. Some may adopt by redoubling their commitment to the procedural epistemic virtues, like carefulness, diligence, and thoroughness, which manifest an admirable commitment to what James Montmarquet called *epistemic conscientiousness* (Montmarquet, 1993: 21). Some people, when confronted with the prospect of deepfakes, may respond by developing epistemic vices: one may abandon any concern to determine the veracity of viewed video content and so exhibit *epistemic insouciance*, a failure to properly care about epistemic goods and values (Cassam, 2018: ch.4). Others may err in another way by radically overestimating their ability to detect deepfakes, assuming their possession of perceptual or discriminative abilities of a type beyond anything they actually possess – a mark of epistemic arrogance or hubris (Tanesini, 2021; Baird and Calvard, 2019: §5; Kraemer, 2014: 25; Kidd, 2018: 59–60).

A more complicated possibility is that the existence of deepfakes radically changes the normative status of epistemic character traits, meaning that epistemic character traits usually regarded as epistemic *vices* under hospitable social-epistemic conditions actually, and rather counterintuitively, function as epistemic *virtues* in hostile environments (Battaly, 2018, 2021). This is the doctrine of *normative contextualism*, whose main claim is that the normative status of epistemic character traits can depend on contextual features of an agent’s socio-epistemic circumstances (Kidd, 2020; Monypenny, 2021). Heather Battaly argues that the trait of closed-mindedness can, in epistemically hostile conditions – ones thoroughly polluted with false beliefs, fake news, and distorting ideologies – function as an epistemic virtue by protecting the few true beliefs one has and keeping out damaging false beliefs (Battaly, 2021).

Perhaps a radical effect of the prevalence of deepfakes is that they alter what sorts of traits should count among what Shannon Vallor calls the *technomoral virtues* (Vallor, 2016). If deepfakes do radically change the nature of our technological socio-epistemic systems, then they might also change the range and status of the epistemic character traits needed to navigate those systems. The radical normative contextualist option is that which traits count as virtues and vices can be changed by the increasing presence of deepfakes. A less radical option may be that deepfakes affect the presence and priority given to different epistemic virtues among the technomoral virtues. Maybe carefulness and diligence become more important, or maybe diffidence and restraint become more important, or perhaps deepfakes create new roles for old-fashioned virtues, like reticence. Whatever option one chooses, it seems clear that a new age of deepfakes will place new pressures on our epistemic characters and capacities, making the already difficult job of responsible epistemic agency even more so.

## 27.4 Conclusion

The aim of this chapter has been to provide an overview of recent and currently unexplored problems that deepfakes raise for epistemologists. We began by discussing two prominent worries articulated by Don Fallis (2020) and Regina Rini (2020): that deepfakes could deflate the evidential role of videos; and that deepfakes could undermine the epistemic trust we place in the testimony and actions depicted in videos.

After that, we addressed three currently unexplored issues that we believe deepfakes will magnify for both *analytic* and *regulative* epistemology (Wolterstorff, 1996). The first was the propensity for deepfakes to introduce *veritic epistemic luck* into our perceptual beliefs from videos, thereby generating new and potential sceptical conclusions. The second was how deepfakes could violate our basic epistemic rights not to be deceived and to know (Watson, 2021). The third problem was the negative effect deepfakes could have on the development of epistemic character, virtue, and vice. While the concerns raised here may seem primarily epistemic, it should be clear that they are indelibly practical, moral, and political, too. That being so, deepfakes offer a rich area of research for those working across both moral and political philosophy.

## Acknowledgements

We are grateful to the editors for their kind invitation to contribute and for their comments on a draft version of this chapter.

## Notes

1. This video is available at: https://www.youtube.com/watch?v=RT2dFzzIedo (2020).
2. A demonstration of this technology available at: https://www.youtube.com/watch?v=RB7upq8nzIU
3. At the time of writing this chapter, the journal *Synthese* issued a call for papers on the philosophical issues around deepfakes.
4. For variations and discussion of safety in epistemology, see Luper (1984), Sainsbury (1997), Sosa (1999), Williamson (2000), and Pritchard (2005).
5. More recently, Pritchard (2015, 2016) has suggested that we ought to focus our attention on what he calls “epistemic risk”, which turns not simply on the non-obtaining of true belief but on the explicit formation of *false beliefs*.
6. This example is adapted from Chisholm (1977).
7. According to recent empirical evidence, a substantial minority of philosophers and non-philosophers are willing to say that you *do* know that the barn you are looking at is genuine, despite the nearby façades. See, for example, Bergenholtz et al. (2021), Calçao et al. (2014) and Turri et al. (2019).
8. This is not to rule out the possibility of deepfakes generating evidential luck. How exactly this might occur is an interesting question, but for purposes of space we cannot entertain it here.
9. Mirsky et al. (2019: 3, 13) found that, of the seventy deepfake scans analysed, the radiologists misdiagnosed 99% with malign tumours, and when informed of the deepfakes they still diagnosed 94% with false tumours.
10. To watch this deepfake, visit https://www.youtube.com/watch?v=IvY-Abd2FfM.

## References

Adjer, H., Patrini, G., Cavalli, F. and Cullen, L. (2019) “The State of Deepfakes”, *Deeptrace*, 1–27, available at: https://regmedia.co.uk/2019/10/08/deepfake\_report.pdf

Atencia-Linares, A. and Artiga, M. (2022) “Deepfakes, Shallow Epistemic Graves: On the Epistemic Robustness of Photography and Videos in the Era of Deepfakes”, *Synthese*, 200(6): 1–22.

Baehr, J. (2010) “Epistemic Malevolence”, *Metaphilosophy*, 41(1–2): 189–213.

——— (2011) *The Inquiring Mind: On Intellectual Virtues and Virtue Epistemology* (Oxford: Oxford University Press).

Baird, C. and Calvard, T.S. (2019) “Epistemic Vices in Organisations: Knowledge, Truth, and Unethical Conduct”, *Journal of Business Ethics*, 160(1): 263–276.

Battaly, H. (2016) “Epistemic Virtue and Vice: Reliabilism, Responsibilism, and Personalism”, in C. Mi, M. Slote, E. Sosa (eds.), *Moral and Intellectual Virtues in Chinese and Western Philosophy: The Turn Towards Virtue* (New York: Routledge), 99–120.

——— (2018) “Closed-Mindedness and Dogmatism,” *Episteme*, 15(3): 261–282.

——— (2021) “Engaging Closed-Mindedly With Your Polluted Media Feed”, in M. Hannon, J. de Ridder (eds.), *The Routledge Handbook of Political Epistemology* (New York: Routledge).

Bergenholtz, C., Busch, J. and Praёm, S. (2021) “Further Insights on Fake Barn Cases and Intuition Variation”, *Episteme*, 20(1): 163–180.

Cassam, Q. (2018) *Vices of the Mind: From the Intellectual to the Political* (Oxford: Oxford University Press).

Cavedon-Taylor, D. (2013) “Photographically based Knowledge”, *Episteme*, 10(3): 283–297.

Chesney, R and Citron, D. (2019) “Deepfakes and the New Disinformation War”, *Foreign Affairs*, 98(1), available at: https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new- disinformation-war

Chisholm, R. M. (1977) *Theory of Knowledge*, 2nd ed. (New Jersey: Englewood Cliffs).

Cohen, J. and Meskin, A. (2004) “On the Epistemic Value of Photographs”, *The Journal of Aesthetics and Art Criticism*, 62(2): 197–210.

Colaço, D., Buckwalter, W., Stich, S. and Machery, E. (2014) “Epistemic Intuitions in Fake-Barn Thought Experiments”, *Episteme*, 11(2): 199–212.

Çolak, B. (2021) “Legal Issues of Deepfakes”, *Institute for Internet & Just Society*, 19.

de Ruiter, A. (2021) “The Distinct Wrong of Deepfakes”, *Philosophy and Technology*, 34(4): 1311–1332.

Dhillon, S. (2019) “An Optimistic View of Deepfakes”, *TechCrunch*, July 4, available at: https://techcrunch.com/2019/07/04/an-optimistic-view-of-deepfakes

Diakopoulos, N. and Johnson, D. (2021) “Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections”, *New Media and Society*, 23(7): 2072–2098.

Engel Jr, Mylan (1992) “Is Epistemic Luck Compatible with Knowledge?”, *Southern Journal of Philosophy*, 30(20): 59–75.

Fallis, D. (2020) “The Epistemic Threat of Deepfakes”, *Philosophy and Technology*, 34(4): 623–643.

Farid, H. (2018) “Digital Forensics in a Post-Truth Age”, *Forensic Science International*, 289: 268–269.

Galson, W. A. (2020) “Is Seeing Still Believing? The Deepfake Challenge to Truth in Politics”, *The Brookings Institute*, 8 January.

Gelfert, A. (2018) “Fake News: A Definition”, *Informal Logic*, 38(1): 84–117.

Gieseke, A. P. (2020) “The New Weapon of Choice: Law’s Current Inability to Properly Address Deepfake Pornography”, *Vanderbilt Law Review*, 73(5): 1479–1516.

Goldman, A. I. (1976) “Discrimination and Perceptual Knowledge”, *Journal of Philosophy*, 73: 771–791.

Groh, M., Epstein, Z., Firestone, C. and Picard, R. (2022) “Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds”, *Proceedings of the National Academy of Sciences*, 119(1): e2110013119.

Habgoode-Coote, J. (2023) “Deepfakes and the Epistemic Apocalypse”, *Synthese*, 201(103): 1–23.

Harris, K. R. (2021) “Video on Demand: What Deepfakes Do and How They Harm”, *Synthese*, 199(5–6): 13373–13391.

Hopkins, R. (2012) “Factive Pictorial Experience: What’s Special About Photographs?”, *Noûs*, 46(4): 709–731.

Jin, Z, Mysore, G. J., Diverdi, S., Lu, J. and Finkelstein, A. (2017) “VoCo: Text-Based Insertion and Replacement in Audio Narration”, *ACM Transactions on Graphics*, 36(4): 96.

Kidd, I. J. (2018) “Deep Epistemic Vices”, *Journal of Philosophical Research*, 43: 43–67.

——— (2020) “Epistemic Corruption and Social Oppression”, in I. J. Kidd, H. Battaly, Q. Cassam (eds.), *Vice Epistemology* (New York: Routledge).

Kidd, I. J., Battaly, H. and Cassam, Q. (2020) *Vice Epistemology* (New York: Routledge).

Kirchengast, T. (2020) “Deepfakes and Image Manipulation: Criminalisation and Control”, *Information and Communications Technology Law*, 29(3): 308–323.

Kraemer, E. (2014) “Virtue Epistemologies and Epistemic Vice”, *Athens Journal of Humanities and Arts* 2(1): 19–30.

Lang, F (2019) “Adobe Trains AI to Detect Deepfakes and Photoshopped Images”, *Interesting Engineering*, available at: https://interestingengineering.com/innovation/adobe-trains-ai-to-detect- deepfakes-and-photoshopped-images

Leonard, N. 2021) “Epistemological Problems of Testimony”, in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, available at: https://plato.stanford.edu/entries/testimony-episprob/

Luper, S. (1984) “The Epistemic Predicament: Knowledge, Nozickian Tracking, and Scepticism”, *Australasian Journal of Philosophy*, 62(1): 26–50.

Lynch, M. P. (2016) *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data* (New York: W.W. Norton and Co).

Matthews, T. (2022) “Deepfakes, Intellectual Cynics, and the Cultivation of Digital Sensibility”, *Royal Institute of Philosophy Supplements*, 91: 67–85.

——— (2023) “Deepfakes, Fake Barns, and Knowledge from Videos”, *Synthese*, 201(41): 1–18.

Mirsky, Y. and Lee, W. (2020) The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, 54(1), 1–41.

Mirsky, Y., Mahler, T., Shelef, I. and Elovici, Y. (2019) “CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning”, *28th USENIX Security Symposium*, available at: https://arxiv.org/pdf/1901.03597.pdf

Montmarquet, J. (1993) *Epistemic Virtue and Doxastic Responsibility* (Lanham, MD: Rowman &

Littlefield).

Monypenny, A. (2021) “Between Vulnerability and Resilience: A contextualist Picture of Protective Epistemic Character Traits”, *Journal of Philosophy of Education*, 55(2): 358–370.

Öhman, C (2022) “The Identification Game: Deepfakes and the Epistemic Limits of Identity”, *Synthese*, 200(319). https://doi.org/10.1007/s11229-022-03798-5

Pfefferkorn, R. (2020) “Deepfakes in the Courtroom”, *Boston University Public Interest Law Journal*, 29(2): 245–276.

Pritchard, D. (2005) *Epistemic Luck* (Oxford, Oxford University Press).

——— (2015) “Risk”, *Metaphilosophy*, 46: 436–461.

——— (2016) “Epistemic Risk”, *Journal of Philosophy*, 113: 550–571.

Rini, R. (2017) “Fake News and Partisan Epistemology”, *Kennedy Institute of Ethics Journal*, 27(2): 43–64.

——— (2020) “Deepfakes and the Epistemic Backstop”, *Philosopher’s Imprint*, 20: 24.

Rini, R. and Cohen, L. (2022) “Deepfakes, Deep Harms”, *Journal of Ethics and Social Philosophy*, 22(2): 141–161.

Sainsbury, R. M. (1997) “Easy Possibilities”, *Philosophy and Phenomenological Research*, 57(4): 907–919.

Sargeant, A. (2017) “The Undeath of Cinema,” *The New Atlantis*, Number 53, (Summer/Fall), 17–32.

Skyrms, B (2010) *Signals: Evolution, Learning, and Information* (Oxford: Oxford University Press).

Sosa, E. (1999) “How to Defeat Opposition to Moore”, *Philosophical Perspectives*, 13: 141–154.

Tanesini, A. (2021) *The Mismeasure of the Self: A Study in Vice Epistemology* (Oxford: Oxford University Press).

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. and Ortega-Garcia, J. (2020) “Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection”, *Information Fusion*, 64: 131–148.

Turri, J., Alfano, M. and Greco, J. (2019) “Virtue Epistemology”, E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, available at: https://plato.stanford.edu/entries/epistemology-virtue/ Vallor, S. (2016) *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford: Oxford University Press).

Watson, L. (2021) *Epistemic Rights and Why We Need Them* (London: Routledge).

Westerlund, M. (2019) “The Emergence of Deepfake Technology: A Review”, *Technology Innovation Management Review*, 9(11): 39–52.

Williamson, T. (2000) *Knowledge and its Limits* (Oxford: Oxford University Press).

WIPO (2019) “WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI)”, *World Intellectual Property Organisation*, December, 2019, available at: https://www.wipo.int/ edocs/mdocs/mdocs/en/wipo\_ip\_ai\_2\_ge\_20/wipo\_ip\_ai\_2\_ge\_20\_1.pdf

Wolterstorff, N. (1996) *John Locke and the Ethics of Belief* (Cambridge: Cambridge University Press).

Zagzebski, L. (1996) *Virtues of the Mind: An Inquiry into the Nature of Knowledge and the Ethical Foundations of Knowledge* (Cambridge: Cambridge University Press).

### *Further resources*

https://www.youtube.com/watch?v=IvY-Abd2FfM – Channel 4’s deepfake of Queen Elizabeth II.

https://www.cbsnews.com/news/artificial-intelligence-holocaust-remembrance-60-minutes- 2020-04-03/. – Storyfile’s deep-learning interviews of Holocaust survivors.

https://thispersondoesnotexist.com/ – a website that uses deep-learning to manufacture entirely non-existent faces.

https://www.youtube.com/watch?v=cQ54GDm1eL0 – the popular deepfake of President Barack Obama